GO analysis walkthrough

By Mary Allen, Jesse Kurland

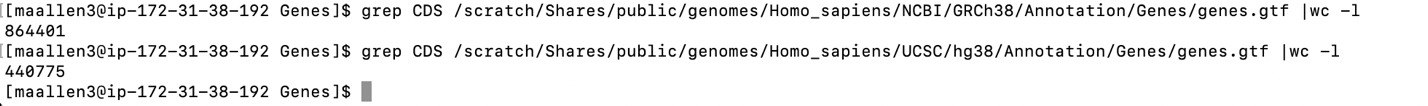
# How does GO term enrichment work?

* scRNA-seq on mouse skeletal muscle
  + Compare to mm10 genome? -> “muscle”
  + Compare to all genes expressed in dataset? -> Identifies different myogenic populations
* Example:
  + Aged vs Adult sRNA-seq from mouse muscle -> 1000 differentially expressed genes in Aged mice
    - **In background gene set**
      * 100,000 total genes in mm10
      * 100 genes involved in innervation of skeletal muscle
    - **In differentially expressed gene set**
      * 100 genes involved in innervation 🡪 not significant!
      * 200 genes involved in innervation 🡪 **significant**!

# Gathering the gene lists

### Before you run Deseq2

Decide on which GTF you will use because some gtfs have more genes than others

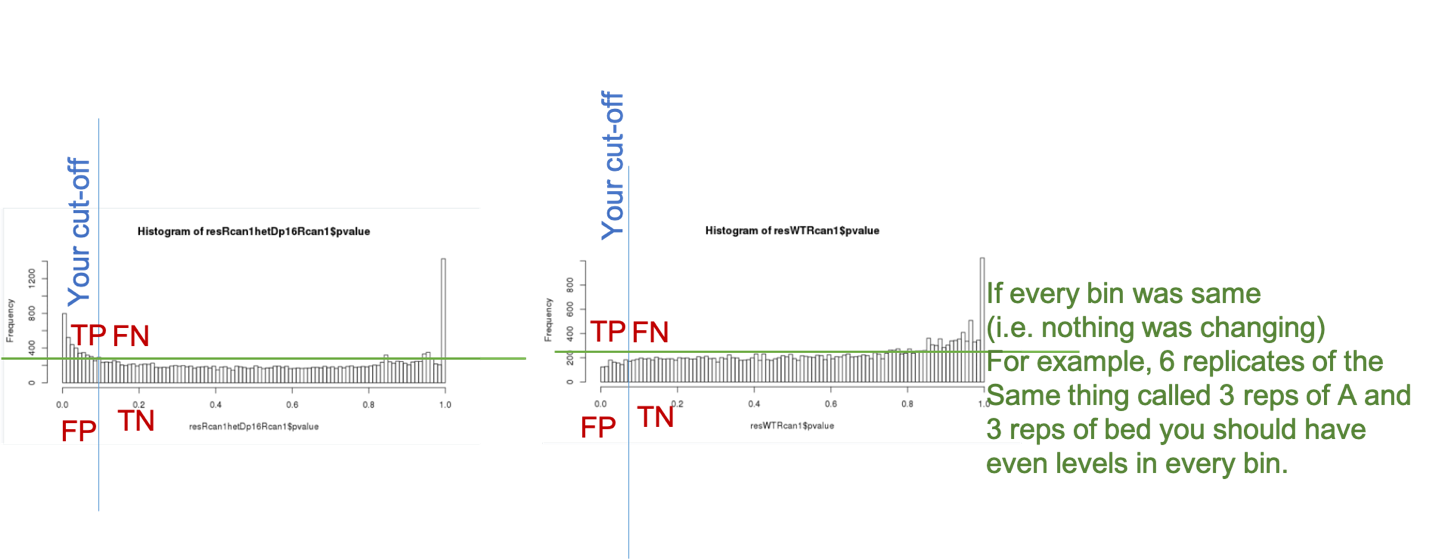


Pro for NCBI/Ensable gtfs: they have way more non-coding RNAs

Con for NCBI/Ensable gtfs: they have way more non-coding RNAs, which means more multiple hypothesis correction and therefore less significant differentaily expressed genes.

### How do I pick my Deseq2 cuttoff?





Draw a histogram of your res$pvalue

Image a blue line at your cut-off and a green line that goes flat across the bins.

These two lines help you to think about your True Positives, False Postives, True Negatives, and False negatives. If you reduce your cut off you get less genes as significant, but more of them are true positives and less of them are false positives.

### How do get my gene lists out of R from Deseq2?

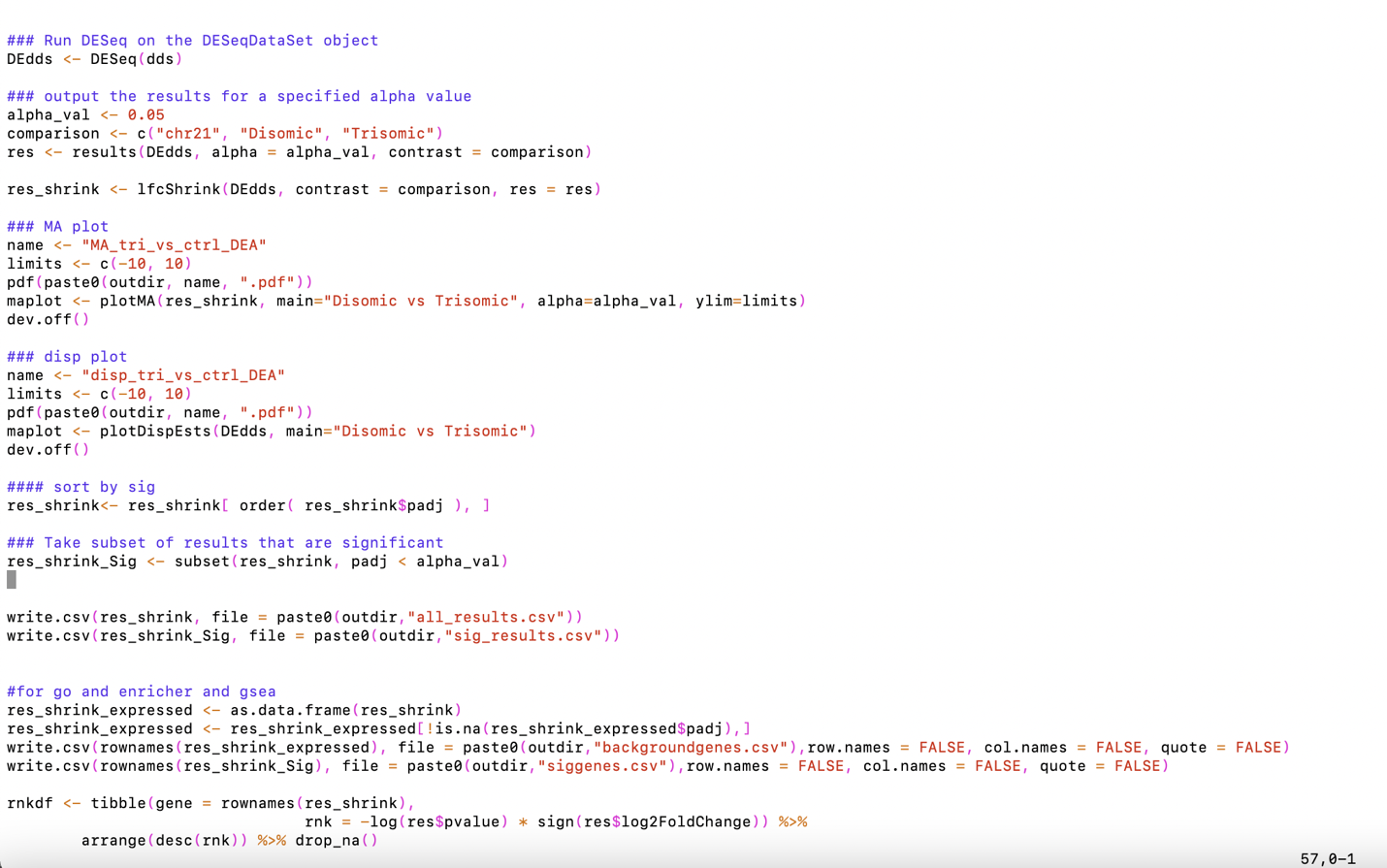
To run GO you will need a significantly different genes list and a background gene list.

* Background gene lists? Which one?

If you could not have called it as differentially expressed it should not be in your background gene list.

The last few lines of this script gather your background gene list and you significant gene list.

Genes that are two low or variable to test for differential expression get a NA in the padj column.



**Enrichr (maayanlab.cloud/Enrichr/)**

* Conducts multiple comparisons (doesn’t permit using background gene set)
* Great for first pass checks of what you should explore more… not the most statically accurate (since not using real background lists)

Graphical user interface, text, application, email

Description automatically generated

* Paste enriched gene list into box and “submit” Graphical user interface, text, application

  Description automatically generated
* “submit” A picture containing calendar

  Description automatically generated

Table

Description automatically generated

**Panther (http://geneontology.org)**

* Allows using background gene sets
* Provides “Molecular Pathways”

Graphical user interface, website

Description automatically generated

“Launch”

Graphical user interface, text, application

Description automatically generated



* Make sure background gene set is in a text file

Graphical user interface, text, application, email

Description automatically generated



Graphical user interface, text, application, email

Description automatically generated



Graphical user interface, application, Word

Description automatically generated